Analysis of variation in regulatory regions using 3,000 rice genomes project

Tatiana V Tatarinova

University of Southern California, Los Angeles, USA, tatarino@usc.edu

Nickolai Alexandrov

International Rice Research Institute, Los Banos, Philippines, n.alexandrov@irri.org

Rice is the staple food for half the world population, particularly for poor developing countries in Asia. Remarkably, rice has a significant within-species genetic diversity. Traditional rice varieties encompass a huge range of potentially valuable genes. These can be used to develop superior varieties for farmers to take part in the uphill battle of feeding an ever-increasing world population (estimated to reach 9.6 billion by 2050). The genes linked to valuable traits can help breeders create new rice varieties that have improved yield potential, higher nutritional quality, better ability to grow in problem soils, and improved tolerance of pests, diseases, and the stresses, such as flood and drought, that will be inevitable with future climate change. Much of this diversity is conserved within the International Rice Genebank Collection (IRGC) at the International Rice Research Institute (IRRI). In the framework of the 3,000 rice genomes project, IRRI and collaborators have completed the sequencing of 3,000 rice genomes of varieties and lines representing 89 countries (Figure 1). The 3,000 Rice Genomes Project Rice Genomes Project is funded by the Bill and Melinda Gates Foundation and the Chinese Ministry of Science and Technology. The project's entire 13.4-terabyte dataset was released in 2014 in an open-access database, GigaDB, which instantly quadrupled the previous amount of publicly available rice sequence data [1]. The dataset contains genome sequences (averaging 14X depth of coverage) derived from 3,000 accessions of rice with global representation of genetic and functional diversity.

Availability of 3K rice genomes provided a unique opportunity to explore variability of different functional regions of genome. We focused our analysis on those regions that are most likely enriched by transcription factor binding sites, such as promoters, 5' and 3'-UTRs. We have examined distribution of SNPs, known transcription factor binding sites, and DNA methylation in those regions. We observe increased sequence conservation in these regions and hypothesize that unusually conserved motifs in these regions have biological significance. We

found the most conserved motifs and performed an enrichment analysis for these motifs in various biological processes.

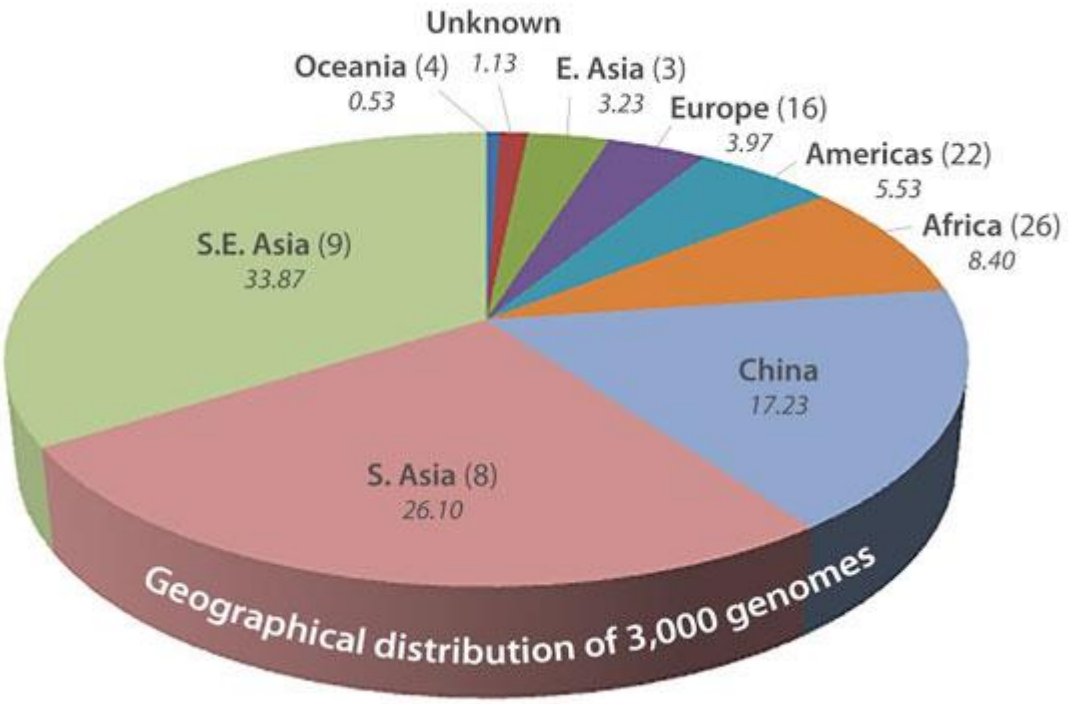

Figure 1: Geographical distribution of 3,000 rice genomes

We applied our reAdmix [2] tool to analysis of 3000 rice genomes, using currently sequenced varieties of wild rice as a reference. We present a novel *plantMix* pipeline for analysis of domesticated species using their wild relatives.

1. Li, J.Y., J. Wang, and R.S. Zeigler, *The 3,000 rice genomes project: new opportunities and challenges for future rice research.* Gigascience, 2014. **3**: p. 8.
2. Kozlov, K., et al., *Differential Evolution Approach to Detect Recent Admixture.* BMC Genomics, 2015.